

Web Archiving Activities of ODU's
Web Science and Digital Library Research Group
@WebSciDL

Michael L. Nelson
@phonedude_mln
Michele C. Weigle
@weiglemc

National Symposium on Web Archiving
Interoperability
2017-02-21

Many projects joint with LANL
Funding from NSF, IMLS, NEH, and AMF

Tools

(warning! research software! YMMV, etc.)

MemGator

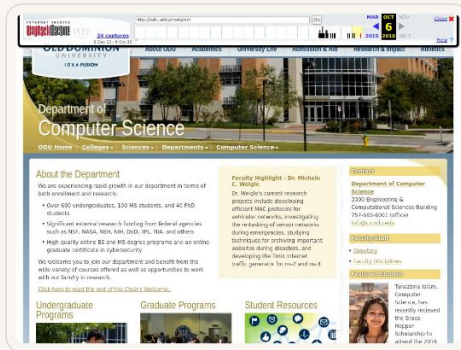
A Memento Aggregator CLI and Server in [Go](#).

Features

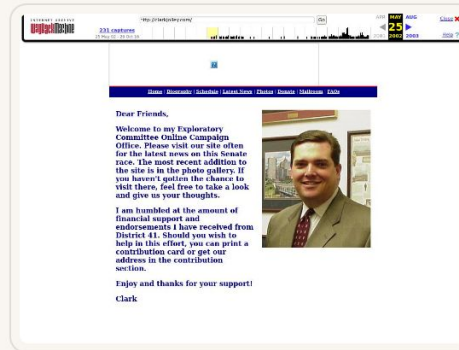
- The binary (available for various platforms) can be used as the CLI or run as a Web Service
- Results available in three formats – Link/JSON/CDXJ
- TimeMap, TimeGate, and Memento (redirect or description) endpoints
- Optional streaming of benchmarks over [Server-Sent Events](#) (SSE) for realtime visualization and monitoring
- Good API parity with the [main Memento Aggregator service](#)
- Concurrent – Splits every session in subtasks for parallel execution
- Parallel – Utilizes all the available CPUs
- Custom archive list (a local JSON file or a remote URL) – a sample JSON is included in the repository
- Probability based archive prioritization and limit
- Three levels of customizable timeouts for greater control over remote requests
- Customizable logging and profiling in CDXJ format
- Customizable endpoint URLs – helpful in load-balancing
- Customizable User-Agent to be sent to each archive and User-Agent spoofing
- Configurable archive failure detection and automatic hibernation
- [CORS](#) support to make it easy to use it from JavaScript clients
- Memento count exposed in the header that can be retrieved via `HEAD` request
- [Docker](#) friendly – An image available as [ibnesayeed/memgator](#)
- Sensible defaults – Batteries included, but replaceable

Sawood Alam (@ibnesayeed) <https://github.com/oduwsdl/memgator>

How well is your webpage archived?



<http://odu.edu/compsci>
Damage = 0.024



<http://clarkjolley.com>
Damage = 0.44014



<http://alguard.state.al.us>
Damage = 0.96800

Check the damage of your page

Check »

Note : The memento damage calculation will work on live webpages, but was designed to evaluate archived webpages or mementos. Discover mementos using [Time Travel](#)

Erika Siregar (@erikaris) <http://memento-damage.cs.odu.edu/>



95 mementos available.

Select a Memento

View

Archive Page To...

List mementos by:

Dropdown

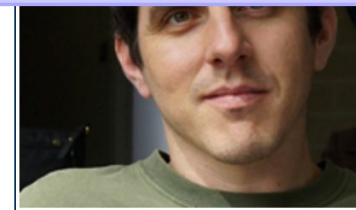
Drilldown

options

Old Dominion University

Current/Upcoming Projects

- Research and paper development on various aspects of personal web archiving.
- Continued development of [WARCreate](#), [Web Archiving Integration Layer \(WAIL\)](#), and [Mink](#) for [NEH Digital Humanities Implementation Grant](#).
- Continued development of [InterPlanetary Wayback \(ipwb\)](#).



[Upcoming/Current Projects](#)
[Papers, Posters & Presentations](#)
[Recognition & Participation](#)
[Teaching](#)
[Coursework History](#)
[Recent Software Projects](#)
[Contact](#)

Papers, Posters & Presentations

[All](#) [Only Peer-Reviewed](#) [Only Papers](#) [Only Journals](#)

- **Mat Kelly**, Sawood Alam, Michael L. Nelson, and Michele C. Weigle. "InterPlanetary Wayback: Peer-To-Peer Permanence of Web Archives," In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)*. Hannover, Germany, September 2016, pp. 411-416. ([PDF](#), [BibTeX](#))
- **Mat Kelly**, Sawood Alam, Michael L. Nelson, and Michele C. Weigle, "InterPlanetary Wayback: The Permanent Web Archive," At the *Web Archiving and Digital Libraries Workshop (WADL 2016)*. Newark, NJ, June 2016.
- Sawood Alam, **Mat Kelly**, and Michael L. Nelson, "InterPlanetary Wayback: The Permanent Web Archive," In *Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. Newark, NJ, June 2016, pp. 273-274. ([PDF](#), [BibTeX](#))
- **Mat Kelly**, "Exploring Aggregation of Personal, Private, and Institutional Web Archives," Presented At *Archives Unleashed 2.0: Web Archive Datathon*, 2016 June 15. ([PPTX](#))
- **Mat Kelly**, "A Framework for Aggregating Private and Public Web Archives," *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, Vol. 11, No. 3, December 2015. ([PDF](#), [BibTeX](#))
- **Mat Kelly**, "A Framework for Aggregating Private and Public Web Archives," at the *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Doctoral Consortium. Knoxville, TN, June 2015. ([PDF](#), [BibTeX](#))
- Wesley Jordan, **Mat Kelly**, Justin F. Brunelle, Laura Vobrak, Michele C. Weigle and Michael L. Nelson, "Mobile Mink: Merging Mobile and Desktop Archived Webs," In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. Knoxville, TN, June 2015, pp. 243-244, **Best Poster Award**. ([PDF](#), [BibTeX](#))
- **Mat Kelly**, Michael L. Nelson, and Michele C. Weigle, "Visualizing Digital Collections of Web Archives," *Web Archiving Collaboration: New Tools and Models*; 2015 June 4; New York City, NY. ([PPTX](#))
- Justin F. Brunelle, **Mat Kelly**, Hany SalahEldeen, Michele C. Weigle and Michael L. Nelson, "Not All Mementos Are Created Equal: Measuring the Impact of Missing Resources," *International Journal of Digital Libraries (IJD)*, 16(3), pp. 283-301. May 2015. ([article](#), [BibTeX](#))
- **Mat Kelly**, "Facilitation of the A Posteriori Replication of Web Published Satellite Imagery", *Virginia Space Grant Consortium 2015 Student Research*
- Justin F. Brunelle, **Mat Kelly**, "The Impact of Missing Resources on the Replication of Web Published Satellite Imagery", *International Journal of Digital Libraries (IJD)*, 16(3), pp. 283-301. May 2015. ([article](#), [BibTeX](#))

[C.V.](#)

Mat Kelly (@machawk1) <https://github.com/machawk1/Mink>

Archive Now (archivenow)

A Tool To Push Web Resources Into Web Archives

Archive Now (archivenow) currently is configured to push resources into four public web archives. You can easily add more archives by writing a new archive handler (e.g., `ia_handler.py`) and place it inside the folder "handlers".

As explained below, this library can be used through:

- CLI
- A Web Service
- A Docker Container
- Python

Installing

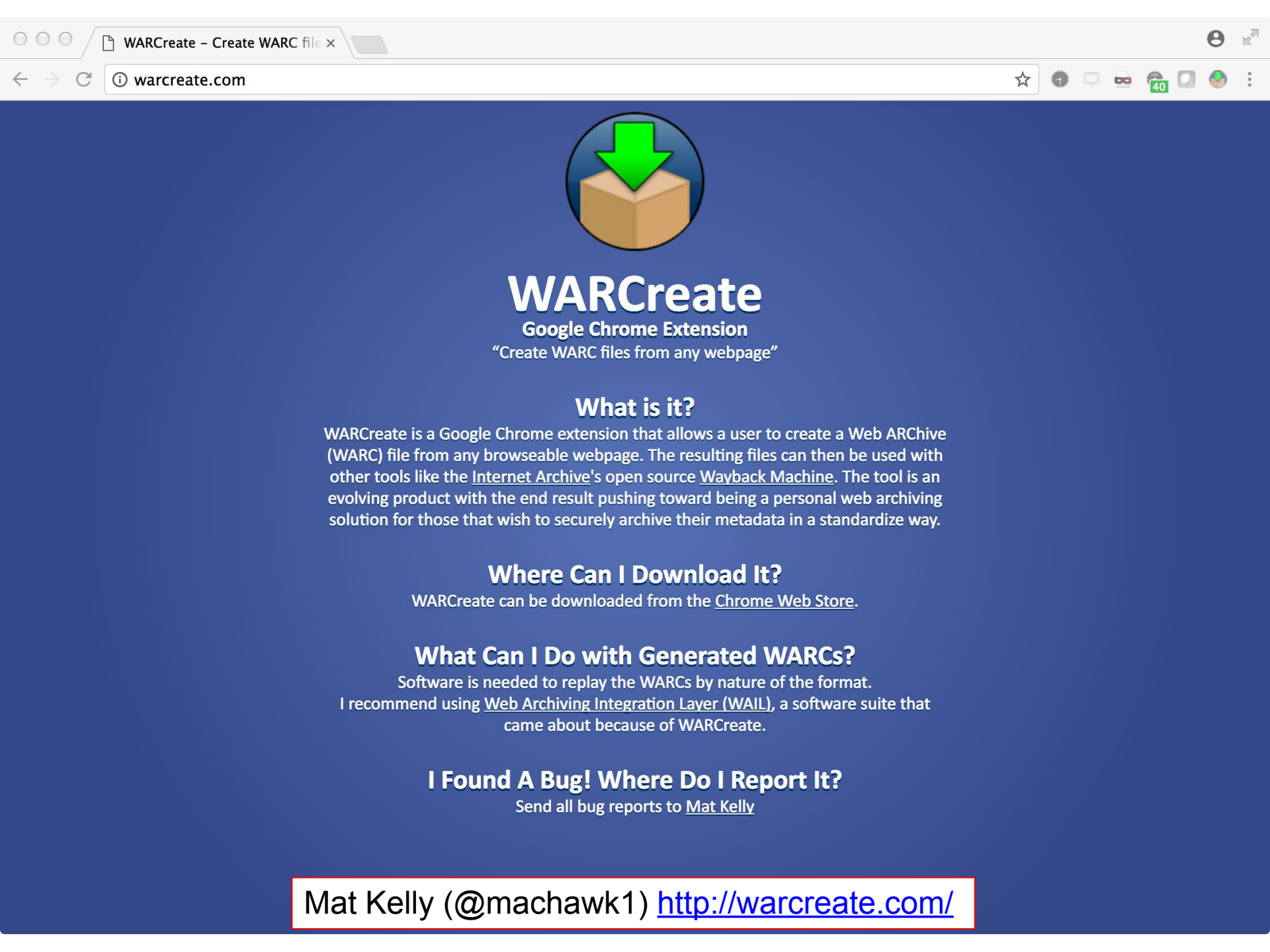
The latest release of archivenow can be installed using pip:

```
$ pip install archivenow
```

The latest development version containing changes not yet released can be installed from source:

```
$ git clone git@github.com:maturban/archivenow.git
$ cd archivenow
$ pip install -r requirements.txt
$ pip install ./
```

Mohamed Aturban (@maturban1) <https://github.com/oduwsdl/archivenow>



WARCreate

Google Chrome Extension

"Create WARC files from any webpage"

What is it?

WARCreate is a Google Chrome extension that allows a user to create a Web ARChive (WARC) file from any browseable webpage. The resulting files can then be used with other tools like the [Internet Archive's](#) open source [Wayback Machine](#). The tool is an evolving product with the end result pushing toward being a personal web archiving solution for those that wish to securely archive their metadata in a standardize way.

Where Can I Download It?

WARCreate can be downloaded from the [Chrome Web Store](#).

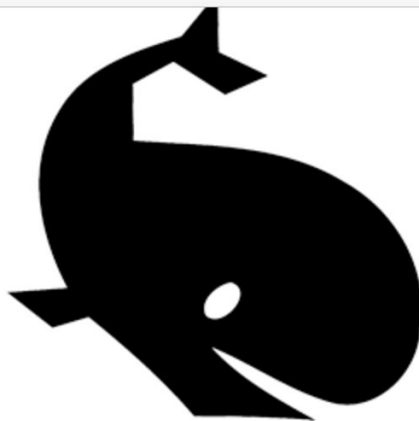
What Can I Do with Generated WARCs?

Software is needed to replay the WARCs by nature of the format. I recommend using [Web Archiving Integration Layer \(WAIL\)](#), a software suite that came about because of WARCreate.

I Found A Bug! Where Do I Report It?

Send all bug reports to [Mat Kelly](#)

Mat Kelly (@machawk1) <http://warcreate.com/>



Web Archiving Integration Layer (WAIL)

"One-Click User Instigated Preservation"

Web Archiving Integration Layer (WAIL)

"One-Click User Instigated Preservation"

Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages. Tools included and accessible through the GUI are [Heritrix 3.2.0](#) and [PyWb 0.33.0](#).

More information about the motivations behind WAIL see the [Motivations](#) section in the projects wiki.

This work is supported by the [National Endowment for the Humanities](#) (NEH), through Digital Humanities grants [HD-51670-13](#) and [HK-50181-14](#)

Wail Electron



John Berlin (@johnaberlin) <https://github.com/N0taN3rd/wail>

Local Memory Project - Chrome x


Secure https://chrome.google.com/webstore/detail/local-memory-project/khineekpnogfcholchjihimhofilcfp

chrome web store

rhodewarriors@gmail.com

Search the store


Featured

 **Local Memory Project**
offered by [localmemoryproject](#)
★★★★★ (0) | [Search Tools](#) | 7 users

+ ADD TO CHROME

OVERVIEW REVIEWS RELATED

G+1 0



LOCAL MEMORY PROJECT


[Get API](#) [Source code](#) [Submit media source](#)

Country: USA Query: protesters and police Zip code: 23529 Media count: 10

Submit


Stories for query: "protesters and police", captured: Tue, Sep 13, 2016, 12:53 PM EDT, for 23529 (Norfolk VA, USA). Archive: [search](#)

- Hampton Roads Messenger, Newspaper, (2.98 miles, VA - USA) [Exclude \(saved/archived\)](#)
- Inside Business, Newspaper, (2.98 miles, VA - USA) [Exclude \(saved/archived\)](#)




Portsmouth council meeting punctuated by hymns, praise, tension ... (Jul 12, 2016)

[Exclude \(saved/archived\)](#)




A moment of healing in Portsmouth: A black protester hugging a white ... (Jul 13, 2016)

[Exclude \(saved/archived\)](#)




Gathering outside the Scope arena in Norfolk protests police shootings (Jul 9, 2016)

[Exclude \(saved/archived\)](#)




Local Black Lives Matter leaders, police praise one another after ... (Jul 11, 2016)

[Exclude \(saved/archived\)](#)




Protesters of police shootings organize in Portsmouth, take to the ... (Jul 11, 2016)


[Exclude \(saved/archived\)](#)




Protesters remained peaceful, ...



Protesters of police shootings hug




Rise: Milwaukee Protesters Throw



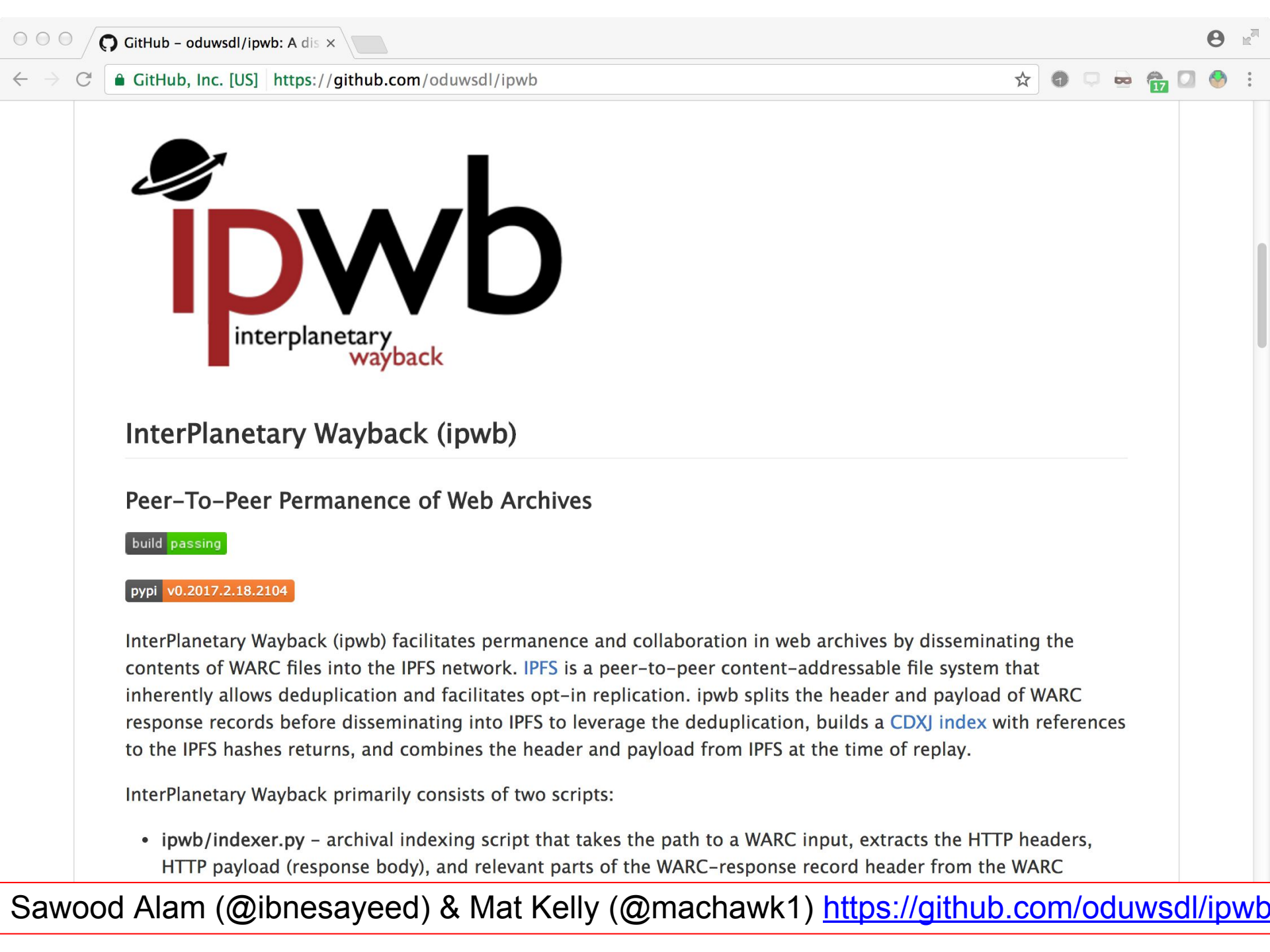
[Police, protesters at DNC praised for restraint, courtesy - PilotOnline.com Jul 25, 2016](#)

[Exclude \(saved/archived\)](#)



Protest against police-involved

Alexander Nwala (@acnwala) <http://www.localmemory.org/> (joint with Harvard)



InterPlanetary Wayback (ipwb)

Peer-To-Peer Permanence of Web Archives

build **passing**

pypi **v0.2017.2.18.2104**

InterPlanetary Wayback (ipwb) facilitates permanence and collaboration in web archives by disseminating the contents of WARC files into the IPFS network. [IPFS](#) is a peer-to-peer content-addressable file system that inherently allows deduplication and facilitates opt-in replication. ipwb splits the header and payload of WARC response records before disseminating into IPFS to leverage the deduplication, builds a [CDXJ index](#) with references to the IPFS hashes returns, and combines the header and payload from IPFS at the time of replay.

InterPlanetary Wayback primarily consists of two scripts:

- `ipwb/indexer.py` – archival indexing script that takes the path to a WARC input, extracts the HTTP headers, HTTP payload (response body), and relevant parts of the WARC-response record header from the WARC

Carbon Dating The Web

Predict the Birthday of a Webpage!

www.cs.odu.edu

Carbon Date!

```
{
  "self": "http://cd.cs.odu.edu/cd?url=www.cs.odu.edu",
  "URI": "http://www.cs.odu.edu",
  "Estimated Creation Date": "1997-03-24T17:29:34",
  "Backlinks": "1997-07-02T13:41:36",
  "Last Modified": "",
  "Bitly.com": "2011-08-31T14:05:22",
  "Twitter.com": "2017-02-19T16:21:10",
  "Bing.com": "",
  "Google.com": "2005-05-28T00:00:00",
  "Pubdate tag": "",
  "Archives": [
    [
      "Earliest",
      "1997-03-24T17:29:34"
    ],
    [
      "By_Archive",
      {
        "http://archive.is/19970606105039/http://www.cs.odu.edu/",
        "http://arquivo.pt/wayback/20091223043049/http://www.cs.odu.edu/",
        "http://webcitation.org/query?id=1327284086752784": "2012-03-24T17:29:34",
        "http://web.archive.org/web/19971010201632/http://www.cs.odu.edu/",
        "http://web.archive.bibalex.org:80/web/20010414022512/http://www.cs.odu.edu/"
      }
    ]
  ]
}
```

Zetan Li <http://cd.cs.odu.edu/>



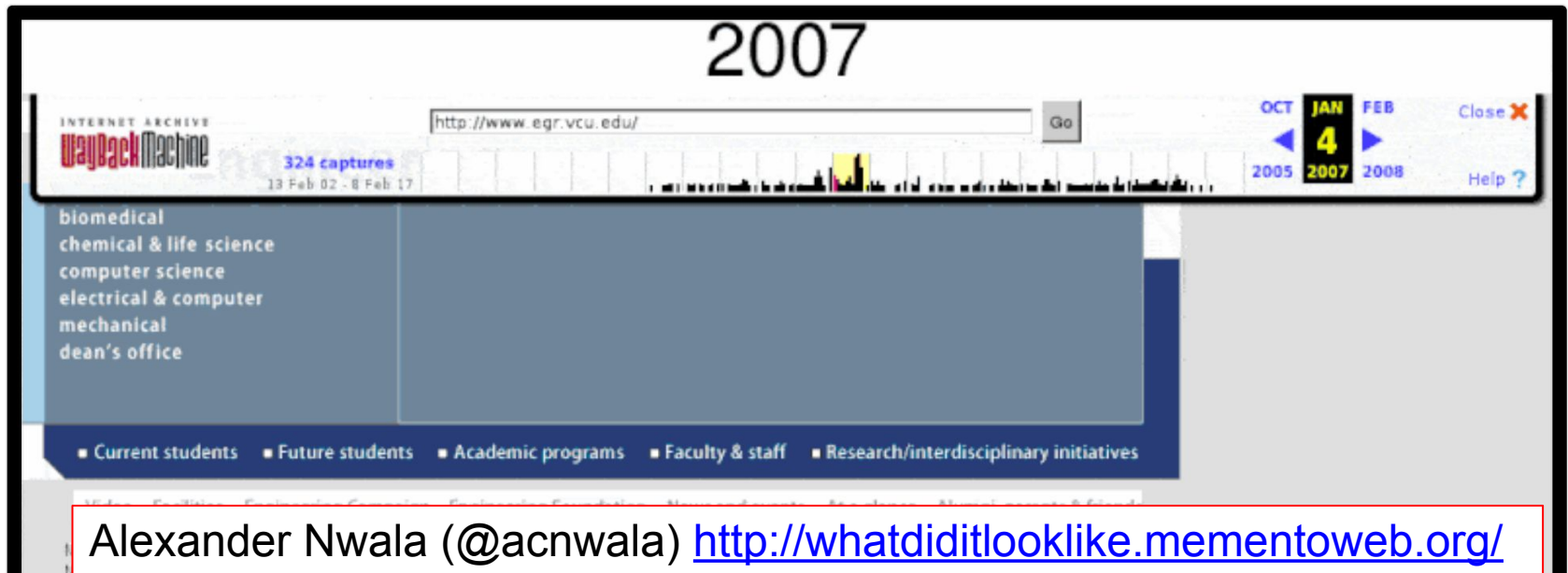
What Did It Look Like ▾

[About](#) [Archive](#) [Follow](#)

What Did It Look Like?

We randomly choose some web sites and see what they looked like through the years.

To nominate a URL for inclusion, tweet "[#whatdiditlooklike](#) URL". Follow [@wdill](#) for daily updates.





Search Twitter



I Can Haz Memento

@icanhazmemento

Use [#icanhazmemento](#) to request links to URLs archived near the time they were shared in your tweet. Via [@WebSciDL](#).

Joined July 2015

[Tweet to I Can Haz Memento](#)

8 Followers you know



TWEETS
164

FOLLOWING
1

FOLLOWERS
27



Following

Tweets

Tweets & replies

In reply to Michael L. Nelson



I Can Haz Memento @icanhazmemento · Jan 25

[@phonedude_mln](#), Your newly archived page:
timetravel.mementoweb.org/memento/201701... (Internet Archive ...).
Other versions: timetravel.mementoweb.org/list/201701250...



In reply to Michael L. Nelson



I Can Haz Memento @icanhazmemento · Jan 24

[@phonedude_mln](#), Your newly archived page:
timetravel.mementoweb.org/memento/201701... (VM Brasseur on Tw...). Other versions: timetravel.mementoweb.org/list/201701250...



In reply to Michael L. Nelson



I Can Haz Memento @icanhazmemento · Jan 24

Wh Alexander Nwala (@acnwala) <https://twitter.com/icanhazmemento>

To detect the off-topic for a collection

- Using collection id from Archive-It

```
python detect_off_topic.py -i [collection_id]
```

For example:

```
python detect_off_topic.py -i 1860
```

- Using collection uri from Archive-It

```
python detect_off_topic.py -r [collection_uri]
```

For example:

```
python detect_off_topic.py -r https://www.archive-it.org/collections/1860
```

- To check off-topic for one timemap

```
python detect_off_topic.py -t [timemap_uri]
```

For example:

```
python detect_off_topic.py -t https://wayback.archive-it.org/2358/timemap/link/http://hamdeensabahy.com/
```

- The default will list the off-topic mementos on the screen, if you want to forward the result to another file

```
python detect_off_topic.py -i [collection_id] -o [filename]
```

- To change the threshold value

```
python detect_off_topic.py -i [collection_id] -th 0.2
```

Shortest Possible PhD Topic Summaries

(+ 1 link for more info)

Temporal Violations, Archive Profiling, Cold Spots

- Scott Ainsworth (@Galsondor)
 - Detecting temporal violations in archival replay
 - <http://ws-dl.blogspot.com/2015/12/2015-12-08-evaluating-temporal.html>
- Sawood Alam (@ibnesayeed)
 - Profiling web archives
 - <http://dx.doi.org/10.1007/s00799-016-0184-4>
- Lulwah Alkwai (@LulwahMA)
 - Eliminating “cold spots” in web archives
 - <http://www.cs.odu.edu/~mln/pubs/jcdl-2015/jcdl-2015-arabic-sites.pdf>

Tampering, Storytelling, Private Web Archives

- Mohamed Aturban (@maturban1)
 - Detecting archival tampering
 - (will have demo at CNI Spring 2017)
- Shawn M. Jones (@shawnmjones)
 - likely to continue Yasmin AlNoamany's storytelling work
 - <https://github.com/yasmina85>
- Mat Kelly (@machawk1)
 - Integrating private and public web archives
 - <http://ws-dl.blogspot.com/2012/08/2012-08-20-ms-thesis-extensible.html>

Automating Archival Collections, Page Transformation, Finding Alumni

- Alexander Nwala (@acnwala)
 - Bootstrapping collections via Twitter, Storify, Reddit, et al.
 - <http://ws-dl.blogspot.com/2016/07/2016-07-18-tweet-visibility-dynamics-in.html>
- John Berlin (@johnaberlin)
 - “It became necessary to destroy the page to save it”
 - (currently MS, will enter PhD)
 - <http://ws-dl.blogspot.com/2017/01/2017-01-20-cnncom-has-been-unarchivable.html>
- Corren McCoy (@CorrenMcCoy)
 - Finding university alumni in social media
 - (only one not working explicitly in web archiving!)
 - <http://ws-dl.blogspot.com/2015/11/2015-11-24-twitter-follower-analysis-of.html>

#IA20

#IA20 – East Coast



<http://ws-dl.blogspot.com/2016/11/2016-11-21-ws-dl-celebration-of-ia20.html>
<https://storify.com/michaelnelson/ws-dl-celebration-of-ia20>